Incorporating Topic Sentence on Neural News Headline Generation

Jan **Wira** Gotama Putra[1], Hayato Kobayashi[2,3], Nobuyuki Shimizu[2]

[1]Tokyo Institute of Technology [2]Yahoo Japan Corporation [3]RIKEN AIP

gotama.w.aa@m.titech.ac.jp {hakobaya, nobushim}@yahoo-corp.jp

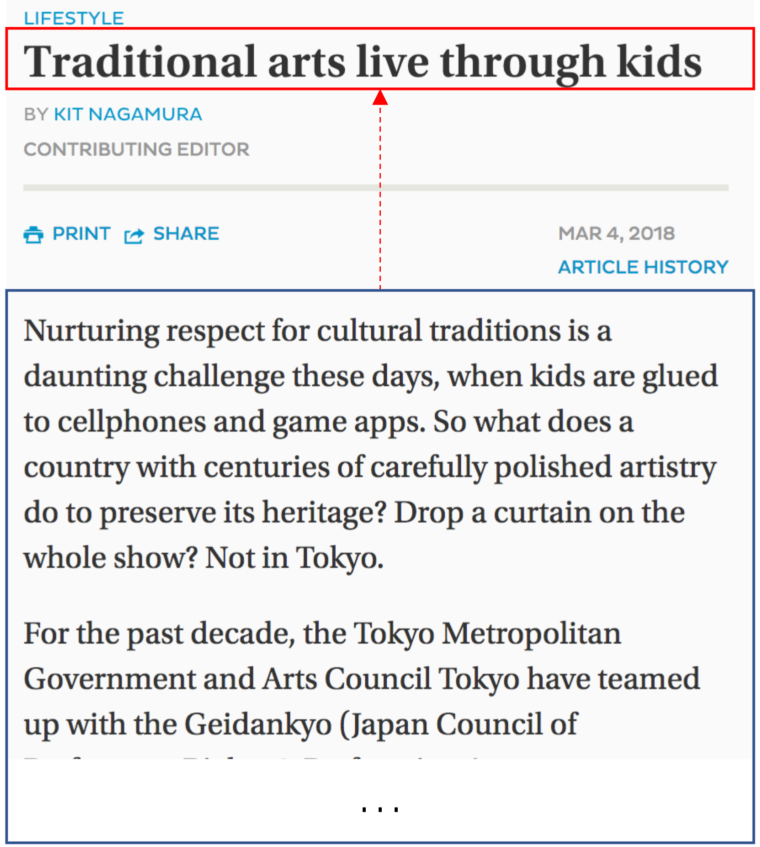
**) Research was done when the first author was an intern (summer 2017) at Yahoo! Japan*

Automatic Headline Generation

- Given a news document, we want to generate a corresponding headline

- Automatic headline generation system is used by news editor as a supporting tool

- Single document summarization
  - *Extractive approach* (Zajic et al., 2004); Colmenares et al., 2015)
  - *Abstractive approach* (Banko, et al., 2000; Rush et al. 2015)


LIFESTYLE

Traditional arts live through kids

BY KIT NAGAMURA

CONTRIBUTING EDITOR

PRINT SHARE

MAR 4, 2018

ARTICLE HISTORY

Nurturing respect for cultural traditions is a daunting challenge these days, when kids are glued to cellphones and game apps. So what does a country with centuries of carefully polished artistry do to preserve its heritage? Drop a curtain on the whole show? Not in Tokyo.

For the past decade, the Tokyo Metropolitan Government and Arts Council Tokyo have teamed up with the Geidankyo (Japan Council of

…

https://www.japantimes.co.jp/life/2018/03/04/lifestyle/traditional-arts-live-kids/#.WqFf8ZOuxsM

Abstractive Headline Generation

- Abstractive approach recently motivated by the success of neural machine translation systems (sequence to sequence) (Sutskever et al., 2014)

- Formalization
  - Given a sequence of N input words (source documents) $\boldsymbol{x} = x_1, x_2, \dots, x_N$
  - The task is to find a sequence of M output words (summary/headline) $\boldsymbol{y} = y_1, y_2, \dots, y_M; M < N$
  - It means we are modeling the conditional probability of input—output pair

    summary $= \arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$

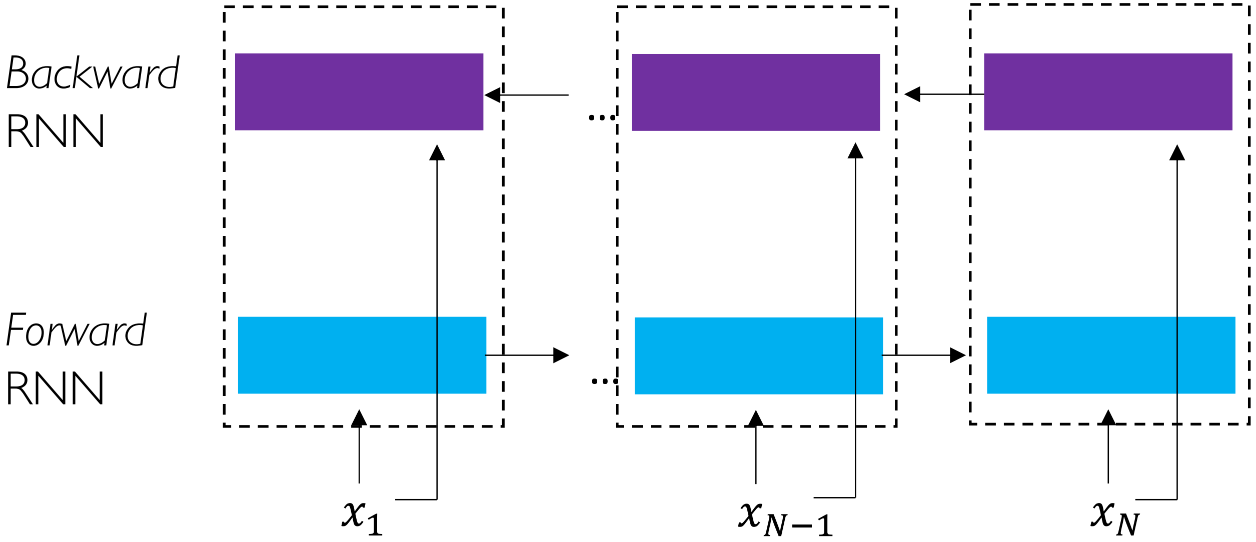
Factoring the Objective

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \prod_{\boldsymbol{t}=\boldsymbol{1}}^{\boldsymbol{M}} \mathrm{P}(y_t|\,\{y_1,\dots,y_t\},\boldsymbol{x},\boldsymbol{\theta})$$

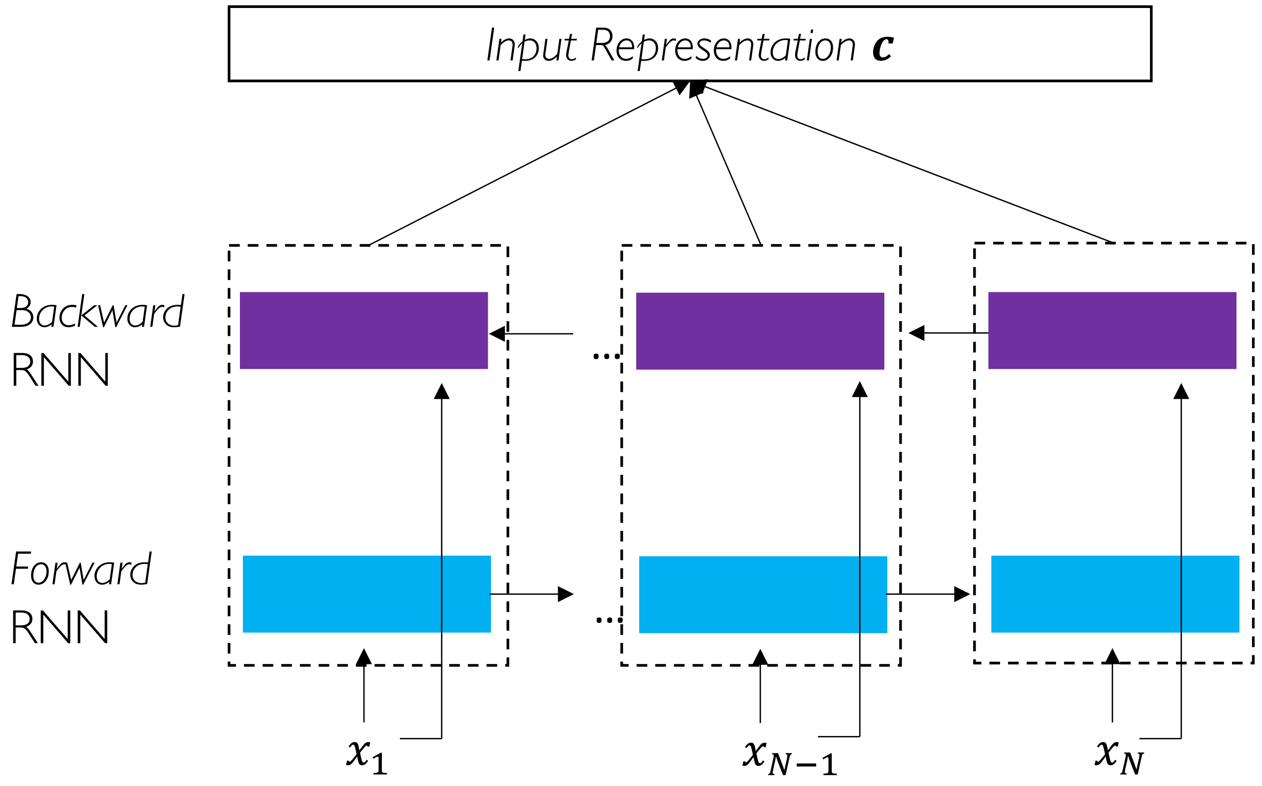
A decoder converts the representation of input ($\boldsymbol{c}$) into a sequence of output $\boldsymbol{y}$

Encoder converts a sequence of input $\boldsymbol{x}$ into a single representation $\boldsymbol{c}$

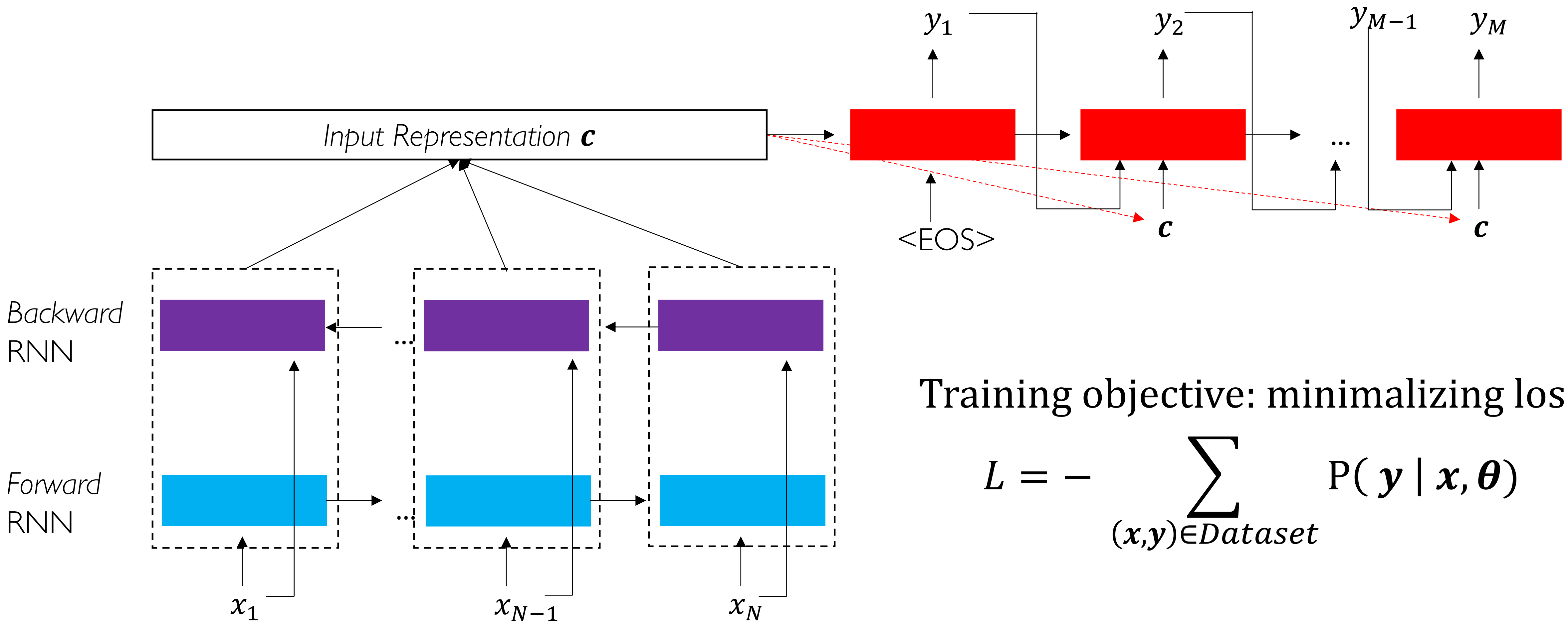
Encoder – Decoder Model



Encoder – Decoder Model



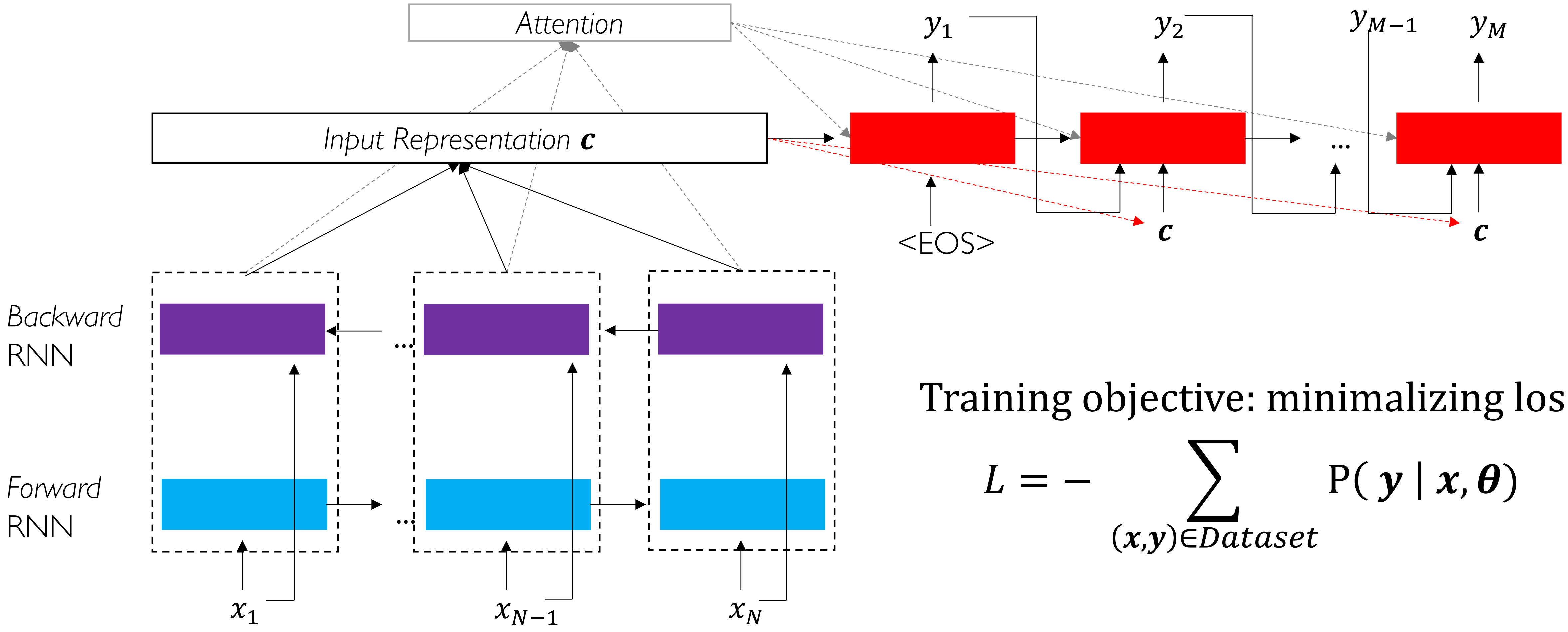
Encoder – Decoder Model



Training objective: minimalizing loss

$$L = - \sum_{(\boldsymbol{x},\boldsymbol{y}) \in Dataset} \mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$$

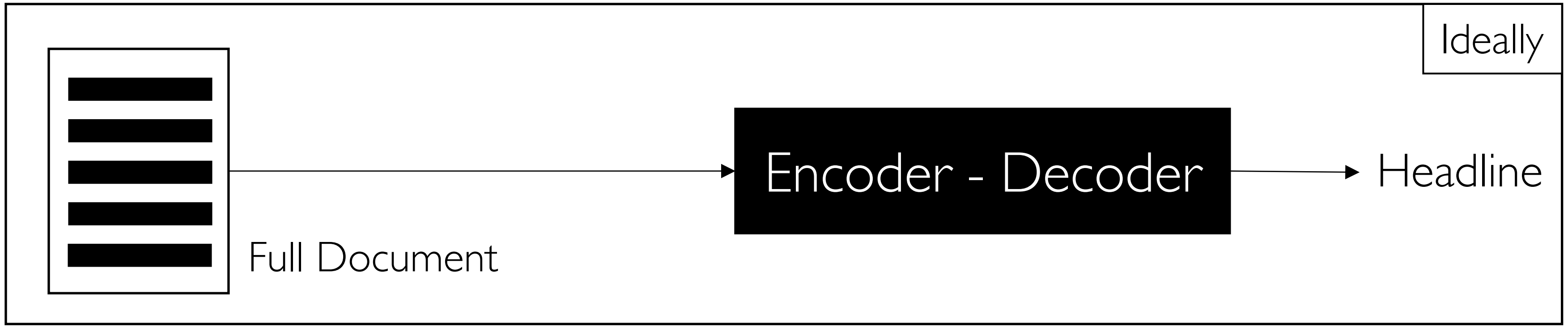
Encoder – Decoder Model with Attention



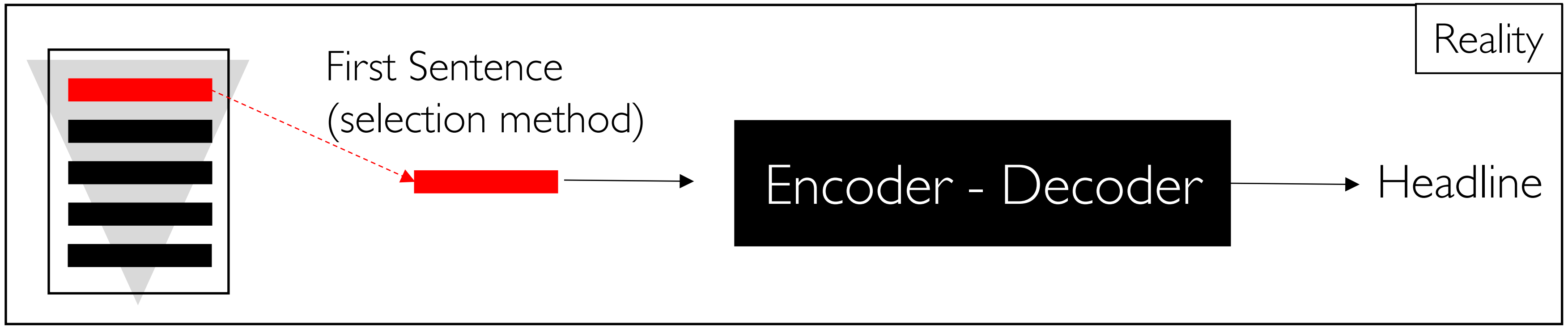
Training objective: minimalizing loss

$$L = -\sum_{(\boldsymbol{x},\boldsymbol{y})\in Dataset} \mathrm{P}(\,\boldsymbol{y}\mid \boldsymbol{x},\boldsymbol{\theta})$$

Related Work



Long Input
Vanishing gradient problem
(Cho et al., 2014; Tan et al., 2017)



Past Studies (headline generation)
Use the first sentence
(Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016; Ayana et al., 2017)

Problems

- The first sentence might not be effective, as the information in a text is distributed across sentences (Alfonseca et al., 2013)
- Using long input may degrade the performance of encoder-decoder (Cho et al., 2014; Tan et al., 2017)
- Previous studies did not consider 5W1H (what, who, when, where, whom, how) information when analyzing news (Wang, 2012).
- How to consider inverse pyramid structure of news (organization structure)

Proposal (contribution)

- Using ***topic sentence* instead of/in addition** to the first sentence

Proposal

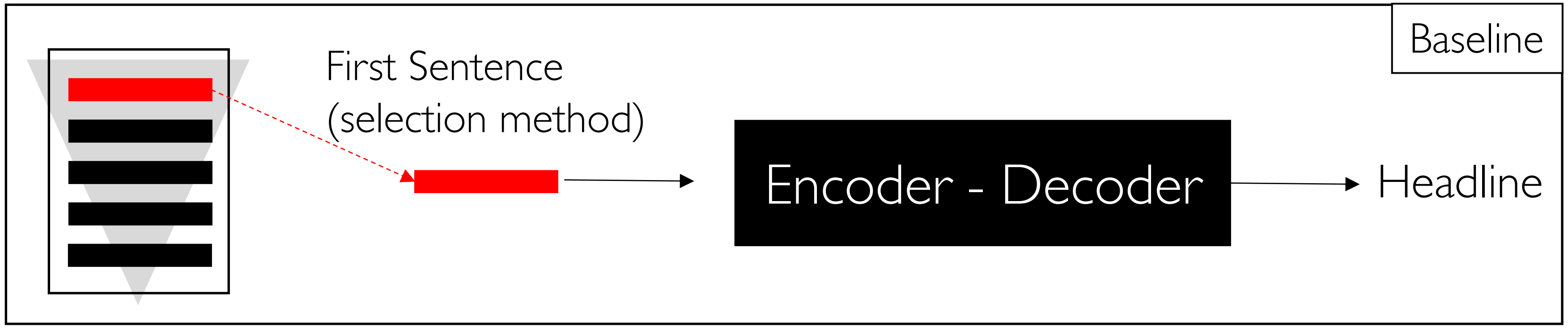
- *Topic sentence* (Wang, 2012) contains **key information of news;** it has the <*subject, verb, object*> elements and at least one subordinate element ***time*** or ***location*** (factual information).
  - *Time* = DATE and TIME (NE tag)
  - *Location* = GPE and LOC (NE tag)

Consider 5W1H (indirectly)

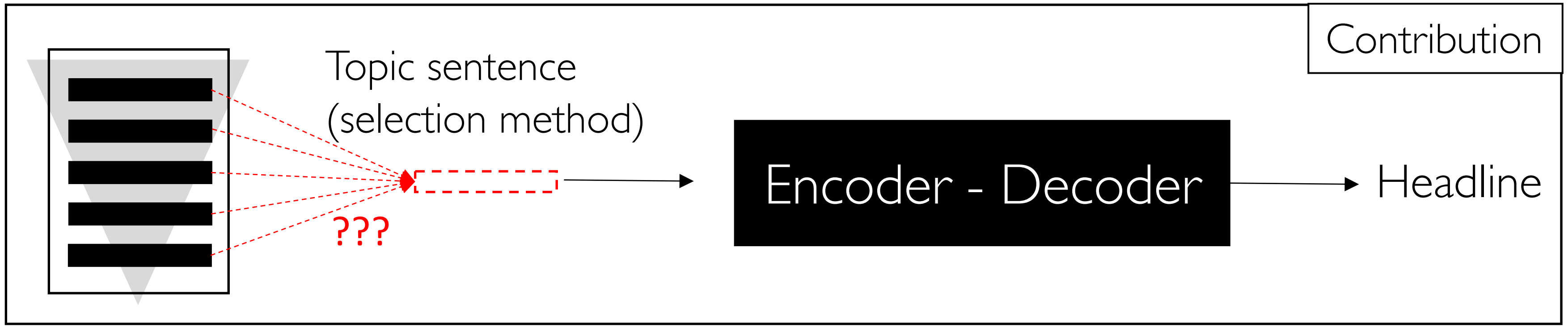
- We extract only one topic sentence from news (the earliest sentence satisfying the rules)

Inverse Pyramid Structure + Short Input

Proposal（contribution）



Past Studies (headline generation)
Use the first sentence
(Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016; Ayana et al., 2017)



Current Study
Use topic sentence for sentence selection

Hypothesis

- We hypothesized that topic sentence is likely to provide a better generalization for the encoder–decoder than using the first sentence

- Generalization means allowing the model to predict the headline of the unseen data in a better way

- *Topic sentence* ≠ *statistical ranking techniques (SRT)*; SRT considers surface information without considering factual information

Experimental Questions

1. Is the topic sentence **more useful** than the first sentence for headline generation?

2. Is the topic sentence **helpful in addition** to the first sentence for headline generation?

Experimental Setting

- We train the encoder—decoder model using three variants of input
  - First sentence (OF)
  - Topic sentence (OT)
  - Both first and topic sentence (OTF)

  and headline (pair)
- We extract only one topic sentence (the earliest sentence satisfying the rules)
- We use the seq2seq implementation of OpenNMT (Klein, et al; 2017)
  - Encoder is 2-layer bidirectional LSTM RNN (500 hidden units)
  - Decoder is 2-layer LSTM RNN (500 hidden units)
  - Global attention mechanism and dropout (0.3) are used

Dataset

- We used Gigaword dataset (10M documents)

Data	# docs	Found-1	Found-2-*	Not found
Train (~90%)	2,755K	2,023K (73.43%)	580K (21.06%)	152K (5.54%)
Valid (~5%)	139K	101K (72.76%)	29K (21.58%)	7K (5.69%)
Test (~5%)	134,K	98K (72.91%)	28K (21.19%)	8K (5.90%)

- Found-1 : Topic sentence is found as the first sentence of the text
- Found-2 : Topic sentence is found as the second or later sentence of the text
- Not found : Topic sentence is not found in the text

Performance

Model	Test Set											
	Topic				First				First and Topic			
	R-1	R-2	R-L	Copy rate	R-1	R-2	R-L	Copy rate	R-1	R-2	R-L	Copy rate
OF	29.45	12.06	26.97	0.72	40.83	20.32	37.97	0.81	23.26	7.90	20.89	0.69
OT	**33.73**	**14.37**	**30.77**	**0.71**	40.71	19.68	37.76	0.80	**26.69**	**8.98**	**23.69**	**0.71**
OTF	32.00	13.03	29.11	0.76	**41.47**	**20.49**	**38.46**	**0.83**	26.49	8.91	23.45	0.75

- *OF* : trained using (first sentence – headline)
- *OT* : trained using (topic sentence – headline)
- *OTF* : trained using (both topic+ first sentences – headline pair)
- R : ROUGE

Output Example

- ***Input:*** for american consumers , the prospect of falling prices sure sounds like a good thing but a prolonged and widespread decline , with everything from real-estate values to income collapsing , would spell disaster for the u.s. economy .
- ***Reference headline:*** falling prices stagnant employment numbers have economists worrying about deflation

- ***OF Prediction:*** u.s. consumer confidence drops to new high
- ***OT Prediction:*** u.s. consumer prices fall #.# percent in may
- ***OTF Prediction:*** u.s. consumer prices fall for first time since ####

Additional Test

Model	Training data	ROUGE		
		R-1	R-2	R-L
OF	2.7 M docs (Rush et al., 2015 + additional filter)	28.38	13.00	26.27
OT		28.77	12.69	26.40
OTF		**29.37**	**13.13**	**27.08**
ABS+	3.7 M docs (Rush et al., 2015)	29.78	11.89	26.97
words-lvt2k-1sent		32.67	15.59	30.64
OpenNMT bechmark*		33.13	16.09	31.00
RAS-Elman		33.78	15.96	31.15
MRT		**36.54**	**16.59**	**31.15**

Small Test Set

2000 first sentence–headline pairs sampled from Gigaword dataset by Rush et al. (2015)

Conclusion

1. Is the topic sentence **more useful** than the first sentence for headline generation?

   Yes, for training (generalization)

2. Is the topic sentence **helpful in addition** to the first sentence for headline generation?

   Yes, it acts as a supporting device

Future Direction

1. Assess the difference of using topic sentence as opposed to other sentence selection/ranking methods

2. Investigate whether using/adding other types of subset of the full news document is able to improve the performance

3. **Automatically decide the optimal subset of text as input for headline generation (encoder-decoder architecture)**

References（1）

1. Zajic, D., Dorr, B. J., and Schwartz, R. (2004). Bbn/umd at duc-2004: Topiary. In *Proceedings of the North America Chapter of the Association for Computational Linguistics Workshop on Document Understanding*, pages 112–119.
2. Colmenares, C. A., Litvak, M., Mantrach, A., and Silvestri, F. (2015). Heads: Headline generation as sequence prediction using an abstract feature-rich space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 133–142.
3. Banko, M., Mittal, V. O., and Witbrock, M. J. (2000). Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*,,pages 318–325.
4. Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.
5. Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3104– 3112.
6. Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning and Representation (ICLR)* .
7. Ayana, Shen, S.-Q., Lin, Y.-K., Tu, C.-C., Zhao, Y., Liu, Z.-Y., and Sun, M.-S. (2017). Recent advances on neural headline generation. *Journal of Computer Science and Technology* , 32(4):768–784, Jul.

References (2)

8. Alfonseca, E., Pighin, D., and Garrido, G. (2013). Heady:News headline abstraction through event pattern clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* , pages 1243–1253.
9. Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* , pages 103–111,
10. Tan, J., Wan, X., and Xiao, J. (2017). From neural sentence summarization to headline generation: A coarse-to-fine approach. *In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4109–4115.
11. Wang, W. (2012). Chinese news event 5w1h semantic elements extraction for event ontology population. In *Proceedings of the 21st International Conference on World Wide Web , WWW '12 Companion*, pages 197–202.
12. Chopra, S., Auli, M., and Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* , pages 93–98.
13. Nallapati, R., Zhou, B., dos Santos, C. N., and C¸ aglar G¨ulc¸ehre and Bing Xiang. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL* .
14. Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL, System Demonstration,* pages 67-72.